CONVERGENCE ANALYSIS OF THE GRAPH ALLEN-CAHN SCHEME[*]

XIYANG LUO[†] AND ANDREA L. BERTOZZI[†]

Abstract. Graph partitioning problems have a wide range of applications in machine learning. This work analyzes convergence conditions for a class of diffuse interface algorithm [A.L. Bertozzi and A. Flenner, *Multiscale Modeling & Simulation*, 10(3):1090 1118, 2012] for binary and multi-class partitioning. Using techniques from numerical PDE and convex optimization, convergence and monotonicity are shown for a class of schemes under a graph-independent timestep restriction. We also analyze the effects of spectral truncation, a common technique used to save computational cost. Convergence of the scheme with spectral truncation is also proved under a timestep restriction inversely proportional to the size of the graph. Moreover, this restriction is shown to be sharp in a worst case. Various numerical experiments are done to compare theoretical results with practical performance.

Key words. diffuse interfaces, convergence, Allen-Cahn equation, Nyström extension

AMS subject classifications. 68U10, 49-04, 49M25, 68T10,

1. Introduction.

Graph cut methods have been widely used in data clustering and image segmentation [9,11,15]. Recently, the reformulation of graph cut problems in graph total variation (TV) minimization has lead to various fast approximations for the optimization [7,8]. In particular, a method inspired by diffuse interface models in PDE was proposed [5]. This approach has been applied to various applications in clustering, image segmentation, and image inpainting [21,23,28].

Classical diffuse interface models are built around the Ginzburg-Landau functional in Euclidean space, defined as

$$GL(u) = \frac{\epsilon}{2}\int |\nabla u|^2 + \frac{1}{\epsilon}\int W(u(x))dx. \tag{1.1}$$

W is the double-well potential $W(u) = \frac{1}{4}(u^2-1)^2$, with minimizers 1 and -1. The first term is the H^1 semi-norm of the function u, which penalizes non-smoothness of u. The parameter ϵ controls the scale of the diffuse interface, namely, the sharpness of the transition between two phases. The Ginzburg-Landau functional and TV is related through the notion of gamma-convergence [30]:

$$\lim_{\epsilon\to 0} GL_\epsilon(u) \to_\gamma CTV(u).$$

Evolution by the Ginzburg-Landau functional has been used to model dynamics of two phases in material science [14,16]. The most common of which is the Allen-Cahn equation, the L^2 gradient flow on the Ginzburg-Landau functional. Under suitable rescaling, the Allen-Cahn equation has been shown to converge to motion by mean curvature [25,36], and thus fast numerical solvers for motion by mean curvature such as the MBO scheme [29] could be utilized to further approximate the Allen-Cahn equation.

The class of methods that we study here are graph analogues of these classical PDE models. Graph Laplacian and graph TV are used in the place of their classical

[*] The work was supported by the UC Lab Fees Research grant 12-LR-236660, NSF grant DMS-1417674, and ONR grant N000141612119

[†] UCLA Department of Mathematics, Los Angeles, CA 90095 (mathluo@math.ucla.edu, bertozzi@math.ucla.edu)

counterparts, and a comprehensive list of these correspondences could be found in [36]. Graph TV minimization and the discrete graph cut problem have been shown to be equivalent in [27]. Moreover, gamma-convergence of graph Ginzburg-Landau to graph TV is proved in [35], justifying its use to approximate the graph cut energy. This paper will focus on the graph Allen-Cahn equation, the L^2 gradient flow of the graph Ginzburg-Landau functional.

Graph Laplacians are themselves of great interest, and have been used in classical machine learning algorithms such as PageRank [12] and spectral clustering [38]. They share many similar characteristics with the continuous Laplacians and can be shown to converge to continuum limits under suitable assumptions [35, 39]. Since they are central to this paper, we introduce below some basics about them below.

Consider a weighted undirected graph G with vertices ordered $\{1, 2, \dots, n\}$. Each pair of vertices (i, j) is assigned a proximity measure w_{ij}. The weights w_{ij} form a weight matrix or adjacency matrix of the graph G. Given a weight matrix W, one can construct three different types of graph Laplacians, namely,

$$L^u = D - W \qquad \text{Unormalized Laplacian,} \tag{1.2}$$

$$L^s = I - D^{-1/2} W D^{-1/2} \qquad \text{Symmetric Laplacian,} \tag{1.3}$$

$$L^{rw} = I - D^{-1} W \qquad \text{Random Walk Laplacian,} \tag{1.4}$$

where D is the diagonal matrix $d_{ii} = \sum_i w_{ij}$. The unnormalized graph Laplacian L^u has the following nice formula for its Dirichlet energy, analogous to that of the continuum Laplacian:

$$\frac{1}{2}\langle u, L^u u\rangle = \frac{1}{2}\sum_{ij} w_{ij}(u_i - u_j)^2. \tag{1.5}$$

The random walk Laplacian L^{rw} has probabilistic interpretations, and can deal with outliers nicely [22]. The symmetric Laplacian L^s shares the same eigenvalues with L^{rw}, but is easier to deal with computationally due to its symmetry. In particular, the unnormalized Laplacian and the normalized Laplacian have very distinct eigenvectors, as can be seen from the visualization in Fig.1.1. The visualization is a plot of the third eigenvector for the nonlocal means graph formed from neighborhood patches of each pixel (see [5] for details). For notational convenience, we omit the superscript for L if the situation applies to all versions of the Laplacians. We discuss all three Laplacians whenever there is a distinction to be made.

In this paper, the only restrictions imposed on the weight matrix W are symmetry and non-negativity. In particular, we do not require triangle inequality. Under this assumption, a useful characterization of an unnormalized Laplacian is given by:

Proposition 1.1 (Characterization of the Unnormalized Graph Laplacian). *Given a matrix L^u, there exists a weight matrix W such that L^u is the corresponding unnormalized graph Laplacian if and only if*

$$\begin{cases} L^u_{ij} = L^u_{ji}, \\ L^u_{ij} \leq 0, i \neq j, \\ L^u_{ii} = -\sum_{j \neq i} L^u_{ij}. \end{cases} \tag{1.6}$$

Proof. See definition of L^u and D. □

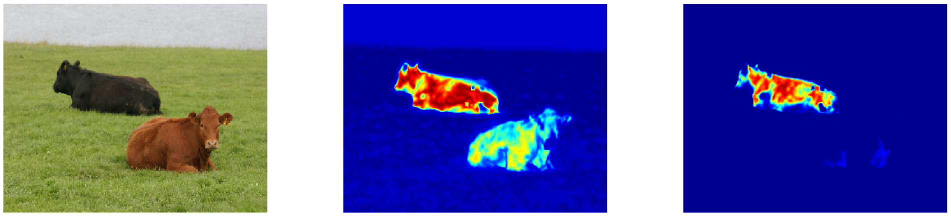

(a) Original Image (b) Symmmetric Laplacian (c) Unnormalized Laplacian

Fig. 1.1: Comparison of Third Eigenvector of Graph Laplacians

We define the Ginzburg-Landau energy on graphs by replacing the spatial Laplacian with the graph Laplacian L.

$$GL(u) = \frac{\epsilon}{2}\langle u, Lu\rangle + \frac{1}{\epsilon}\sum_i W(u_i). \tag{1.7}$$

Note from here on L could be one of three versions of the graph Laplacian . The Allen-Cahn equation on graphs is defined as the gradient flow of the graph Ginzburg-Landau energy

$$u_t = -\nabla GL(u) = -\epsilon Lu - \frac{1}{\epsilon}W'(u). \tag{1.8}$$

In [5], a semi-implicit numerical scheme was used to counter the ill-conditioning of the graph Laplacian

$$\frac{u^{k+1} - u^k}{dt} = -\epsilon Lu^{k+1} - \frac{1}{\epsilon}W'(u^k). \tag{1.9}$$

Moreover, a convex penalty $\frac{c}{2}u^2$ can be used to form a "convex-concave" split, this gives us the actual scheme in [5]

$$\frac{u^{k+1} - u^k}{dt} = -\epsilon Lu^{k+1} - cu^{k+1} + cu^k - \frac{1}{\epsilon}W'(u^k). \tag{1.10}$$

Convex-concave splitting originated in an unpublished paper by Eyre [17], and has been used to resolve long-time solutions of the Cahn-Hilliard equation in [4,37]. These Cahn-Hilliard equations were meant to resolve the time dynamics whereas here we are simply trying to find the equilibrium point. Therefore, for the particular scheme (1.10), we show in the next proposition that it is equivalent with (1.9) and thus we henceforth only analyze the scheme without convex splitting (1.9).

Lemma 1.2 (Rescaled Timestep). *The scheme (1.10) is the same as*

$$u^{k+1} - u^k = -\frac{dt}{1 + cdt}\epsilon Lu^{k+1} - \frac{dt}{1 + cdt}\frac{1}{\epsilon}W'(u^k). \tag{1.11}$$

Hence (1.10) is equivalent to (1.9) under a rescaling of stepsize.

In practice, a fidelity term $f(\phi(u), u)$ is often added. Since the Ginzburg-Landau functional is a smoothed version of TV, this model is reminiscent of the classical ROF model [31] used in imaging when the fidelity term is quadratic. In the next sections, we analyze the scheme (1.9) without fidelity first, and later incorporate the fidelity term in the analysis.

2. Maximum Principle-L^∞ Estimates.

The main result for this section is the following:

PROPOSITION 2.1 (A Priori Boundeness). *Define u^k by*

$$u^k = [(1 + \frac{dt}{\epsilon} L^u)]^{-1} (u^{k-1} - \frac{1}{\epsilon} W'(u^{k-1})), \tag{2.1}$$

derived from (1.9), and L^u is the unnormalized graph Laplacian Assume $\|u^0\|_\infty \leq 1$. If $dt \leq 0.5\epsilon$, then $\forall k$, $\|u^k\|_\infty \leq 1$.

Details of the proposition will be covered in the next two sections. What is notable is that the timestep restriction is independent of the graph, i.e., this universal bound is guaranteed to work for *any* graph of *any* size. The constant $.5\epsilon$ is analogous to the ODE stiffness condition(see [33]). To prove the result, we need the maximum principle on graphs.

2.1. Maximum Principle.

The classical maximum principle argument relies on the fact that $\Delta u(x_0) \geq 0$ for x_0 a local minimizer. This fact is also true for graphs.

PROPOSITION 2.2 (Second Order Condition on Graphs). *Let u be a function defined on a graph, and L^u be the unnormalized graph Laplacian . Suppose u achieves a local minimum at a vertex i, then $L^u u|_i \leq 0$, where a local minimizer i is defined as $u_i \leq u_j, \quad \forall w_{ij} > 0$.*

Proof. Let i be a local minimizer. Then

$$\begin{aligned} L^u u|_i &= d_i u_i - \sum_{j \neq i} w_{ij} u_j \\ &= \sum_{j \neq i} w_{ij}(u_i - u_j) \leq 0. \quad \square \end{aligned} \tag{2.2}$$

Next, we prove a discrete analogue of the continuous time maximum principle, which states that the implicitly discretized scheme for the heat equation on graphs is decreasing in the L_∞ norm. This line of thought is inspired by the maximum principle for finite difference operators [13].

PROPOSITION 2.3 (Maximum Principle for Discrete Time). *For any $dt \geq 0$, let u be a solution to*

$$u = -dt * (Lu) + v, \tag{2.3}$$

then we have $\|u\|_\infty \leq \|v\|_\infty$.

Proof. Suppose i is a maximum of u. Then since $u(i) = dt * (-Lu)(i) + v(i)$ and $(-Lu)(i) \leq 0$, we have $max_u = u(i) \leq v(i) \leq max_v$. Arguing similarly with the minimum, we have that $\|u\|_\infty \leq \|v\|_\infty$. □

Remark: The condition $L_{ij} \leq 0, i \neq j$ is *crucial* for the above analysis to hold. If L is replaced by a general positive-semidefinite matrix, we still have an L^2 version of the proposition, namely, $\|u\|_2 \leq \|v\|_2$, but the L^∞ version is not true.

2.2. Proof of Proposition 2.1.

We use the maximum principle to prove that the sequence $\{u^k\}$ is uniformly bounded under a suitable initial condition. Consider splitting the one-line scheme (1.9) into two parts.

$$\begin{cases} v^k = u^k - dt * \frac{1}{\epsilon} W'(u^k), \\ u^{k+1} = -dt * (\epsilon L^u u^{k+1}) + v^k. \end{cases} \tag{2.4}$$

Note that by the maximum principle, $\|u^{k+1}\|_\infty \leq \|v^k\|_\infty$. We need the lemma below to control the L^∞ norm of the first line in (2.4).

LEMMA 2.4. *Define the map*

$$\mathcal{F}_{dt}(x) = x - dtW'(x) = x - dtx(x^2 - 1). \tag{2.5}$$

If $dt < 0.5$, $\mathcal{F}_{dt}$ maps $[-1, 1]$ to itself.

Proof. Note $\mathcal{F}_{dt}(\pm 1) = \pm 1, \forall dt$, thus $\mathcal{F}_{dt}$ maps endpoints to endpoints. $\mathcal{F}_{dt}$ maps the entire interval to itself if it is monotone on the interval. Computing the roots r_i of $\mathcal{F}'_{dt}(r) = 0$, $r_i = \pm\sqrt{\frac{1}{3}(\frac{1}{dt} + 1)}$. Since $\mathcal{F}_{dt}$ is qubic, it is easy to see that $\mathcal{F}_{dt}$ is monotone on $[-1, 1]$ iff $|r_i| > 1$, i.e., $dt < 0.5$. □

Note that $\mathcal{F}_{dt}(x)$ here is the component-wise map of the first line in (2.4) and is the source of the stepsize restriction. Since estimates for other forward steps $\mathcal{F}_{dt}(x)$ follows the same idea as above and involves only elementary calculations, we omit some of the details later. Next, we prove our main conclusion below:

Proof. [Proposition 2.1] We prove by induction. Suppose $\|u^k\|_\infty \leq 1$. Then for each vertex i, we have $v^k(i) = u^k(i) - \frac{dt}{\epsilon}W'(u^k(i)) = \mathcal{F}_{dt/\epsilon}(u^k(i))$. By Lemma 2.4, $|v^k(i)| < 1, \forall i$, given that $\frac{dt}{\epsilon} < 0.5$. Hence $\|v^k\|_\infty \leq 1$. Then by Proposition 2.3, $\|u^{k+1}\|_\infty \leq \|v^k\|_\infty \leq 1$. □

If we are not so keen on keeping $u^k(i)$ in $[-1, 1]$ for each iteration, but merely care about whether the scheme is bounded or not, then we may relax the interval a bit to get a larger range of dt, as the next lemma shows.

PROPOSITION 2.5. *For $dt < 2.1$, $\mathcal{F}_{dt}$ maps $[-1.4, 1.4]$ to itself. Hence if $\|u^0\|_\infty < 1.4$, then $\{u^k\}$ is bounded for $dt < 2.1\epsilon$.*

Derivation of the constants: Define $\phi(c)$ to be the maximum dt for which $\mathcal{F}_{dt}$ maps $[-c, c]$ to itself. Since $\mathcal{F}_{dt}$ is qubic, ϕ is continuous. The exact constants are then found by using a computer program to brute force maximize ϕ.

For future reference, the 0.5ϵ bound will be called the "tight bound" where the 2.1ϵ bound will be called the "loose bound". Again, we must emphasis that the exact constants here do not matter so much as the fact that they are independent of the graph.

2.3. Generalizations of the scheme. In this section, we prove boundedness results for other versions of the scheme.

PROPOSITION 2.6 (Random walk graph Laplacian). *Let $\|u^0\|_\infty \leq 1$. If $dt < 0.5\epsilon$, the scheme (1.9) with $L = L^{rw}$ satisfies $\|u^k\|_\infty \leq 1, \forall k$.*

Proof. The proof follows similarly from that of Proposition 2.1 by noting that the second order condition (2.2) holds also for random walk Laplacian L^{rw}. □

The case on symmetric graph Laplacian is a little different, and a uniformity condition on the graph must be added.

PROPOSITION 2.7 (Symmetric graph laplacian). *Define the uniformity constant ρ as*

$$\rho = \frac{\max_i d_i}{\min_i d_i}. \tag{2.6}$$

If $\rho \leq 2$, $\|u^0\|_\infty \leq \frac{1}{\sqrt{2}}$, $dt \leq 0.5\epsilon$, then the sequence $\{u^k\}$ is bounded.

Proof. We note that while L^s_{ij} is no longer negative (thus losing the maximum principle), we still have the relation $L_s = D^{1/2}L_{rw}D^{-1/2}$. Thus line 2 of (2.4) becomes

$$D^{-1/2}u^{k+1} = -dt * L_{rw}D^{-1/2}u^{k+1} + D^{-1/2}v^k. \tag{2.7}$$

By the maximum principle, $\|D^{-1/2}u^{k+1}\|_\infty \leq \|D^{-1/2}v^k\|_\infty$. We rescale u^k as $\tilde{u}^k = \sqrt{\min_i d_i} \times D^{-1/2}u^{k+1}$,and $\tilde{v}^k = \sqrt{\min_i d_i} \times D^{-1/2}v^{k+1}$. The first line of (2.4) becomes

$$\tilde{v}^k(i) = \frac{1}{c_i}\mathcal{F}_{dt/\epsilon}(c_i\tilde{u}^k(i)),$$

where $c_i = (\frac{d_i}{\min_j d_j})^{1/2}$. Note that by uniformity condition (2.6), $c_i \in [1, \sqrt{2}], \forall i$.

Next, we prove $\|\tilde{u}^k\|_\infty \leq 1$ by induction. This is clearly true for $k = 0$ since $\|\tilde{u}^0\|_\infty \leq \sqrt{\rho}\|u^0\|_\infty \leq 1$. For general k, define $\Phi_i(x) = \frac{1}{c_i}\mathcal{F}_{dt/\epsilon}(c_i x)$ to be a rescaled version of the forward step. We claim Φ_i maps $[-1, 1]$ to itself for all i. This follows easily from the following lemma:

Lemma 2.8 (Uniform scaling). *For any $dt < 0.5$, $\mathcal{F}_{dt}$ as defined in (2.5) maps $[-c, c]$ to itself for any $c \in [1, \sqrt{2}]$.*

The proof of the Lemma 2.8 can be done using brute force calculation similar to that of Lemma 2.4. Thus we get $\tilde{v}^k(i) \in [-1, 1]$. Applying the maximum principle and equation (2.7), we finally get $\|\tilde{u}^{k+1}\|_\infty \leq 1$ and complete the induction argument. □

In the context of semi-supervised learning [5,21], a quadratic fidelity term appears in the scheme. We show boundedness of the graph Allen-Cahn scheme with this added fidelity. Restating from [5] the scheme with fidelity:

$$\begin{cases} v^k = u^k - dt * (\frac{1}{\epsilon}W'(u^k) + \eta\Lambda(u^k - \phi^0)), \\ u^{k+1} = - dt * (\epsilon L u^{k+1}) + v^k, \end{cases} \tag{2.8}$$

where $\phi^0(i) \in \{1, -1\}$, for i belonging to a fidelity set Λ.

Proposition 2.9 (Graph Allen-Cahn with fidelity). *The graph Allen-Cahn scheme with fidelity (2.8) is bounded by 2 for $dt < \frac{1}{2+\eta}\epsilon$, where η is the fidelity strength and ϵ is the diffuse parameter.*

Proof. Define the modified forward step $\Phi_{dt}(u) = u - dt[(u^2 - 1)u + \eta(u - 1)]$. By the same argument as in Lemma 2.4 by noting Φ is qubic, the limit stepsize is given by $\Phi'_{dt}(1) = 0$. Thus $dt = \frac{1}{2+\eta}\epsilon$. □

2.4. Uniform Convergence to ODE. As an interesting corollary of the Maximum Principle, we prove that the discrete scheme u^k converges to the ODE equation in a graph independent way.

Proposition 2.10 (Convergence to ODE). *Let U be the continuous time solution of (1.8) defined on a fixed interval $[0, T]$, where L is either the unnormalized Laplacian or random walk Laplacian. Assume $\|U(., 0)\|_\infty \leq 1$. Then the semi-implicit scheme in (1.9) has first order converges uniformly to the continuous time solution U with respect to L.*

Proof. Set $U^k = U(., kdt)$ Define $e^k = u^k - U^k$ to be the global error. Then we have to show $\|e_k\| \leq Mdt$, where M is independent of dt and the graph Laplacian L.

By assumption, $\|u^0\|_\infty \leq 1$. Therefore by Proposition 2.1 we have that the sequence $\{u^k\}$ is uniformly bounded. Since U is a gradient flow, $GL(U(., t))$ is decreasing in time and thus U is also bounded. Therefore, we may assume W'' to be bounded (by a graph-free constant). In the following notation, M and the constant for big O is graph-independent. We first compute the local truncation error:

Lemma 2.11. *Define the local truncation error τ^k as below:*

$$\frac{U^{k+1} - U^k}{dt} = -LU^{k+1} - W'(U^k) + \tau^k. \tag{2.9}$$

Then $\|\tau^k\|_\infty \leq Mdt$ Subtract (2.9) from the discrete scheme, and we get an evolution equation for the error term

$$\frac{e^{k+1}-e^k}{dt} = Le^{k+1} - (W'(u^k) - W'(U^k)) + \tau^k. \tag{2.10}$$

By using Taylor expansion on W', we have

$$e^{k+1} = -dtLe^{k+1} - dtW''(\xi)e^k + e^k + \tau^k. \tag{2.11}$$

Here W'' is the diagonal matrix $W''_{jj} = W''(\xi_j)$. By applying the maximum principle, taking L^∞ norm gives us

$$\|e^{k+1}\|_\infty \leq \| - dtW''(\xi)e^k + e^k + \tau^k\|_\infty \leq (1 + Mdt)\|e^k\|_\infty, \tag{2.12}$$

for some graph independent constant M.

Hence by iterating the equation above, we get

$$\|e^k\|_\infty \leq (Me^{MT})dt = O(dt). \quad \square$$

3. Energy method-L^2 estimates. In this section, we derive estimates in terms of the L^2 norm, commonly used to prove convergence of finite difference schemes of parabolic PDEs. Our goal is to prove convergence to stationary point for the full scheme and also convergence of scheme with spectral truncation. Our proof is loosely motivated by the analysis on convex-concave splitting in [17,40]. For example in [17], Eyre proved:

PROPOSITION 3.1 (Eyre). *Let $E = E_1 + E_2$ be a splitting with E_1 convex and E_2 concave. Then for any dt, the semi-implicit scheme $u^{k+1} = u^k - dt\nabla E_1(u^{k+1}) - dt\nabla E_2(u^k)$ is monotone in E, namely,*

$$E(u^{k+1}) \leq E(u^k), \quad \forall k \geq 0.$$

The semi-implicit scheme in fact coincides with the notion of proximal gradient method for minimizing the splitting $E = E_1 + E_2$. Recall that proximal gradient iteration is given as [6]

$$u^{k+1} = Prox_{dtE_1}(u^k - dt\nabla E_2(u^k)), \tag{3.1}$$

where the *Prox* operator is defined as

$$Prox_{\gamma f}(x) = argmin_u\{f(u) + \frac{1}{2\gamma}\|u - x\|^2\}.$$

The *Prox* operator is defined for all proper convex functions taking extended real values, and its connection to the semi-implicit scheme is clear from its implicit gradient interpretation. Namely, if $y = Prox_{\gamma f}(x)$,

$$y = x - \gamma\partial f(y). \tag{3.2}$$

∂f is the *subgradient* of f, which coincides with the gradient if f is differentiable. Classical results for convergence of proximal gradient method can be found in [6],which requires both E_1, E_2 to be convex (instead of E_2 concave as in Eyre), and ∇E_2 be Lipschitz continuous.

In our case, $E = GL(u)$, $E_1 = \frac{\epsilon}{2}\langle u, Lu\rangle$ and $E_2 = \frac{1}{4\epsilon}\sum_i W(u_i)$. However, our E_2 is neither concave nor convex. Instead of using a quadratic penalty to force convex-concavity (which is shown to be a rescaling of time in (1.2)) we follow the analysis for done for non-convex proximal gradient method. The estimate below is a simplified version of the proof in [32], and is analogous to that in [17]. General discussions of non-convex proximal gradients are found in [2].

PROPOSITION 3.2 (Energy Estimate). *Let $E = E_1 + E_2$. Define x^{k+1} by $x^{k+1} \in x^k - dt\partial E_1(x^{k+1}) - dt\nabla E_2(x^k)$. Suppose E_1 is convex, E_2 smooth and ∇E_2 Lipschitz continuous with Lipschitz constant M, we have*

$$E(x^k) - E(x^{k+1}) \geq (\frac{1}{dt} - \frac{M}{2})\|x^{k+1} - x^k\|^2. \tag{3.3}$$

Proof.

$$\begin{aligned} E(x^k) - E(x^{k+1}) &= E_1(x^k) - E_1(x^{k+1}) + E_2(x^k) - E_2(x^{k+1}) \\ &\geq \langle \partial E_1(x^{k+1}), x^k - x^{k+1}\rangle + E_2(x^k) - E_2(x^{k+1}) \\ &= \frac{1}{dt}\|x^{k+1} - x^k\|^2 + \langle \nabla E_2(x^k), x^k - x^{k+1}\rangle + E_2(x^k) - E_2(x^{k+1}) \\ &\geq \frac{1}{dt}\|x^{k+1} - x^k\|^2 - \frac{M}{2}\|x^{k+1} - x^k\|^2. \end{aligned}$$

The second line is by convexity of E_1, the third by plugging in the the definition of x^{k+1}, the fourth by simple Taylor expansion of the function E_2. □

Even though the proof is simple, we have the freedom of choosing E_1 to be a general non-smooth convex function. In particular, we later set $E_1(u) = I_V(u) + \frac{\epsilon}{2}\langle u, Lu\rangle$, where I is the indicator function and V an eigenspace.

3.1. Convergence of Scheme.

In this section, we use the estimate (3.2) to extend our result of boundedness to convergence under the same stepsize restriction.

PROPOSITION 3.3 (Convergence of Graph Allen-Cahn). *Let $\|u^0\|_\infty \leq 1$. Under the strict bound $dt < 0.5\epsilon$, the scheme (1.9) is monotone in the Ginzburg-Landau energy $GL(u) = \frac{\epsilon}{2}\langle u, Lu\rangle + \frac{1}{\epsilon}\sum_i W(u_i)$ and converges to a stationary point of GL.*

Proof. From Proposition 2.1, $dt < 0.5\epsilon$, implies $\|u^k\|_\infty \leq 1, \forall k$. We set $E_1 = \frac{\epsilon}{2}\langle u, Lu\rangle$, $E2 = \frac{1}{\epsilon}\sum_i W(u_i)$, and apply Proposition (3.2). Since the L^∞ ball in $\mathbb{R}^n$ is convex, all points ξ that lie in the line segment from u^k to u^{k+1} satisfy $\|\xi\|_\infty \leq 1$. Thus we can WLOG assume the Lipschitz constant M of ∇E_2 to satisfy $M \leq \frac{1}{\epsilon}\max_{|x|_\infty \leq 1}|W''(x)| = \frac{2}{\epsilon}$. Since $dt < 0.5\epsilon < \frac{2}{M}$, by Proposition 3.2, we have:

$$GL(u^n) - GL(u^{n+1}) \geq (\frac{1}{dt} - \frac{M}{2})\|u^{k+1} - u^k\|^2, \tag{3.4}$$

where $\frac{1}{dt} - \frac{M}{2} > 0$ by our assumption on dt. Hence u^k is monotone in GL. Summing both sides of (3.4) we obtain

$$GL(u^0) - GL(u^n) \geq (\frac{1}{dt} - \frac{M}{2})\sum_{i \leq n}\|u^i - u^{i-1}\|^2. \tag{3.5}$$

Since $GL(u^n) \geq 0$ and $\frac{1}{dt} - \frac{M}{2} > 0$, the sequence $\{u^i - u^{i-1}\}$ is square summable, thus $\lim\limits_{i\to\infty}\|u^i - u^{i-1}\| = 0$. Since $\{u^i\}$ is bounded, there exists a limit point u^* for the sequence, and u^* is unique by the condition $\lim_{i\to\infty} \|u^i - u^{i-1}\| = 0$. Hence $\{u^i\}$ converges to u^*. By continuity, u^* is a stationary point, i.e., $\nabla GL(u^*) = 0$. □

Remark: The argument does not work for the "loose bound" $dt < 2.1\epsilon$, since the Lipschitz restriction $\frac{2}{M}$ is no longer greater than 2.1ϵ. There indeed exists cases where the sequence $\{u^k\}$ is bounded but does not converge. However, in practice the scheme tends to converge under larger timestep than the tight bound 0.5ϵ.

Following the same line of proof, above and using results in Section 2, we obtain a general convergence result for the graph Allen-Cahn scheme.

Theorem 3.4 (Main Convergence Result). *Let u^k be defined by a form of the Ginzburg-Landau scheme as below:*

$$u^{k+1} = u^k - dt * (\epsilon L^* u^{k+1} + \frac{1}{\epsilon} W'(u^k) + \eta\Lambda(u^k - \phi^0)), \tag{3.6}$$

L^* being either the unnormalized or random walk Laplacian. Then if $\|u^0\|_\infty \leq 1$, then $\exists c$ independent of L such that $\forall dt < c$, we have $\lim_{k\to\infty} u^k = u^*$, where u^* is a stationary point of the Ginzburg-Landau functional. The result holds for symmetric Laplacians if we add an additional uniformity condition (2.6) on the graph.

3.2. Convergence Results for General Semi-Definite L. In this section, we study the scheme (1.9) where L is replaced by an *arbitrary* symmetric semi-positive definite matrix. We desire to prove similar convergence results but without the maximum principle. Instead, we rely solely on the energy estimates itself.

There are two reasons for generalizing L to an arbitrary semi-positive definite matrix. First, this serves as a baseline convergence result if we are to study symmetric perturbations made on the original L, such as in the case of using the Nystrom Extension. The second reason is that the proof here can be carried over to show convergence of (1.9) under spectral truncation. Our main result is below:

Theorem 3.5. *Let L be an arbitrary symmetric and semi-positive definite matrix such that $\rho_L \leq C$ for some C independent of N, where $\rho_L = \max_i |\lambda_i|$. Define u^k by the scheme (1.9) with $\epsilon = 1$, i.e.,*

$$\begin{cases} v^k = u^k - dt * W'(u^k), \\ u^{k+1} = -dt * (Lu^{k+1}) + v^k. \end{cases} \tag{3.7}$$

Suppose $\|u^0\|_\infty \leq 1$, then the scheme is monotone in the Ginzburg-Landau energy for timestep $dt = O(N^{-1})$, where N is the size of the system, i.e., number of vertices in the graph.

Since the result is an analysis on dt vs N, we allow the constants to depend on ϵ, and WLOG set $\epsilon = 1$ in the proof.

Our strategy here is to choose dt so small and apply Proposition 3.2 to force *monotonicity* in the Ginzburg-Landau functional GL. Since $GL(u) = O(\|u\|^4)$, boundedness in function value implies boundedness in the variable u. For our purpose, we need a more refined version of the bound on $GL(u)$.

Lemma 3.6 (Inverse Bound). *Let M be any positive constant. If $GL(u) \leq M$, then $\|u\|_2^2 \leq N + \sqrt{NM}$, where N is the dimension of u.*

Proof. By assumption on $GL(u) = \sum_i (u_i^2 - 1)^2 + \langle u, Lu\rangle$, $\sum_i (u_i^2 - 1)^2 \leq M$. Then from the Cauchy-Schwarz inequality, $\sum_i (u_i^2 - 1) \leq \sqrt{NM}$, hence our lemma. □

To prove the theorem, we need a direct estimate on the norm of u^{k+1} from u^k. The proof of the lemma is an application of norm conversions in Lemma 8.1 in the Appendix.

LEMMA 3.7. *Let u^k and u^{k+1} be successive iterates defined in (3.7). Then their function value satisfies the inequality below:*

$$\|u^{k+1}\|_2 \leq (1+dt)\|u^k\|_2 + dt\|u^k\|_2^3. \tag{3.8}$$

Proof. Since L is symmetric semi-positive definite, we have $\|u^{k+1}\|_2 \leq \|v^k\|_2$. Then since $v^k(i) = u^k(i) - dt * [u^k(i)(u^k(i)^2 - 1)]$, we have $\|v^k\|_2 \leq (1+dt)\|u^k\|_2 + dt\|u^k\|_6^3 \leq (1+dt)\|u^k\|_2 + dt\|u^k\|_2^3$ □

We can now prove the main conclusion:

Proof. [Proposition 3.5.] Since $\|u^0\|_\infty \leq 1$, we have $\|u^0\|_2 \leq \sqrt{N}$. Moreover, $GL(u^0) \leq \rho_L\|u^0\|_2^2 + \sum_{i\leq N} 1 \leq C_1 N$. By Lemma 3.6, for any v s.t. $GL(v) \leq GL(u^0)$, we have $\|v\| \leq C_2\sqrt{N}$, for some $C_2 \geq 1$.

We claim that there exists a constant δ independent of N such that $\forall dt \leq \delta N^{-1}$, (3.9) holds for all k. Note that C_1 and C_2 are independent of *both* N and the interation number k.

$$\begin{aligned} GL(u^k) \leq& GL(u^0) \leq C_1 N, \\ \|u^k\|_2 \leq& C_2\sqrt{N}. \end{aligned} \tag{3.9}$$

We argue by induction. The case $k = 0$ has been proved above. Suppose this is valid for k, then we have $\|u^k\| \leq C_2\sqrt{N}$. By Lemma 3.7, we have $\|u^{k+1}\| \leq \frac{A_1}{2}(1+dt)N^{1/2} + \frac{A_1}{2}dtN^{3/2}$, where A_1 depends on $C1, C2$ but not k or N. Therefore, we can choose $dt \leq \delta N^{-1}$, such that $\|u^{k+1}\| \leq A_1 N^{1/2}$. Again, δ is not dependent on k or N.

The crux of the proof is applying the energy estimate (3.2) to u^{k+1}. Note the estimate (3.2) is valid with constant $M \leq \max_{\|\xi\|_\infty \leq A_1\sqrt{N}} \|\nabla^2 W(\xi)\|$. Plugging in the explicit formula $W(u) = \frac{1}{4}(u^2-1)^2$, we have $M \leq A_2 N$ for some A_2. Thus by further shrinking δ if necessary, we have for $dt < \delta N^{-1}$, $\frac{1}{dt} - \frac{M}{2} > 0$. Hence applying (3.2), we have. $GL(u^{k+1}) \leq GL(u^k) \leq GL(u^0)$. However, this would mean that $GL(u^{k+1}) \leq C_1 N$, and thus by the inverse bound Lemma 3.6, $\|u^{k+1}\|_2 \leq C_2\sqrt{N}$. This completes the induction step. □

Remark: The convergence stepsize *is* graph-size dependent. However, we show in the next section that the dependence in unavoidable if we are dealing with arbitrary L.

4. Analysis on Spectral Truncation.

In many applications, the number of nodes N on a graph is huge, and it is infeasible to invert L every iteration in (1.9). In [5, 28], a strategy proposed was to project u onto the first m eigenvectors. This approach is called *spectral truncation.* In practice, spectral truncation gives accurate segmentation results but is computationally much cheaper. There are several methods for precomputing the eigenvectors including Nystrom method [18] which is a random sampling method, and the Raleigh-Chebyshev method [1] for sparse matrices.

4.1. Convergence Results for Spectral Truncation.

Let us formally define the spectral truncated version of scheme (1.9). Define $V_m = span\{\phi^1, \phi^2, \ldots, \phi^m\}$ to be space spanned by the m eigenvectors of L with the smallest eigenvalues. Define P_m to be the projection operator onto the space V_m. Then the spectral truncated scheme is defined as

$$\begin{cases} v^k = u^k - dt * \frac{1}{\epsilon}W'(u^k), \\ u^{k+1} = P_m[-dt * (\epsilon L u^{k+1}) + v^k]. \end{cases} \tag{4.1}$$

Just as in the previous analysis, we first show boundedness.

PROPOSITION 4.1 (Boundedness of Spectral Truncation). *Let the graph Laplacian satisfy $\rho_L \leq C$ for some C independent of N. Define u^k by the scheme (4.1). Suppose $\|u^0\|_\infty \leq 1$, then the scheme is monotone in the Ginzburg-Landau energy for timestep $dt = O(N^{-1})$, where N is the size of the system.*

The key to the proof is the following observation, which links spectral truncation to the proximal gradient method.

LEMMA 4.2 (Reformulation of Spectral Truncation). *If $u^0 \in V_m$, the spectral truncated scheme (4.1) is equivalent to the proximal gradient scheme (3.1) with $E_1 = \frac{\epsilon}{2}\langle u, Lu\rangle + I_{V_m}$, $E_2 = \frac{1}{4\epsilon}\sum_i W(u_i)$, where I_{V_m} is the indicator function of the m-th eigenspace which is 0 inside V_m and $+\infty$ outside.*

Proof. Since the forward step, i.e., line 1 of (4.1) is the same between the two schemes, we only have to show the following: Define u, u' as

$$u = \underset{y}{\operatorname{argmin}} \frac{\epsilon}{2}\langle y, Ly\rangle + \frac{1}{dt}\|y - v\|^2,$$
$$u' = \underset{y}{\operatorname{argmin}} \frac{\epsilon}{2}\langle y, Ly\rangle + I_{V_k}(y) + \frac{1}{dt}\|y - v\|^2.$$

We want to show that u, u' satisfy the relation $P_m(u) = u'$. Project onto eigenvectors and we have $u = \sum_{i=1}^{N} c_i \phi^i$, $u' = \sum_{i=1}^{N} c'_i \phi^i$. Since $u' \in V_m$, we have $c'_i = 0, \forall i > m$. Moreover, letting $d_i = \langle v, \phi^i\rangle$,

$$\frac{\epsilon}{2}\langle u, Lu\rangle + \frac{1}{dt}\|u - v\|^2 = \sum_{i=1}^{N}(\frac{\epsilon}{2}\lambda_i c_i^2 + (c_i - d_i)^2).$$

Thus the minimization is done component-wise in c_i, and it is easy to see that $c_i = c'_i, \forall i \leq m$. Thus $P_m(u) = u$. □

We may now prove our main result:

Proof. [Proposition 4.1] The proof is almost identical to that in Proposition 3.5. We again argue inductively that (3.9) holds for $dt \leq \delta N^{-1}$. Suppose (3.9) is true for k. Since the the projection operator P_k does not increase L^2 norm, we still have $\|u^{k+1}\| \leq (1 + dt)\|u^k\| + dt\|u^k\|^3$, and thus $\|u^{k+1}\| \leq C\sqrt{N}$. By the equivalence of spectral truncation with proximal gradient (4.2), we may use the energy estimate (3.2), and argue that $GL(u^{k+1}) \leq GL(u^k)$ under $dt < \delta N^{-1}$. This in turn forces $\|u^{k+1}\| \leq C_2\sqrt{N}$, ending the induction. □

Since we now know that $\|u^k\|$ is bounded by $O(\sqrt{N})$, we can WLOG assume the Lipschitz constant $M = O(N)$. Thus by following the proof of Proposition 3.3, we have

PROPOSITION 4.3 (Convergence Result). *The truncated scheme is convergent under the stepsize restriction $dt \leq \delta N^{-1}$.*

4.2. A Counter Example for Graph-Independent Timestep Restriction. In the previous subsection, we proved that the spectral truncated scheme is bounded under stepsize restriction $dt = O(N^{-1})$. One would hope to achieve a graph-free stepsize rule as in the case of the full scheme (1.9). However, as we show in our example below, uniform convergence stepsize over all graph Laplacian of all sizes is not possible.

PROPOSITION 4.4 (Optimality of Estimate 4.1). *For any δ independent of N and $dt = \delta N^{-\alpha}, \alpha < 1$, we can always find a graph Laplacian $L_{N\times N}$ with $\rho_L \leq 1$, and an initial condition $\|u^0\|_\infty = 1$ such that the scheme in (4.1) with truncation level $m = 2$ has $\lim\limits_{k\to\infty} \|u^k\|_\infty = \infty$. Hence our estimate for the stepsize restriction in (4.1) is optimal.*

We explicitly construct a graph and a graph Laplacian to attain the worst case bound. Graph construction is as follows, which is illustrated in Fig 4.1.

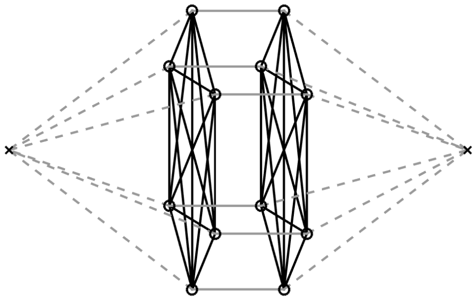


Fig. 4.1: Illustration of Worst Case Graph with $N = 7$

Construction of Graph

1. *Nodes*: The nodes consists of two groups of "clusters" nodes (in circles) and two outlier nodes (in x). Each cluster contains $N - 1$ nodes and thus the graph contains a total of $2N$ nodes.

2. *Edge Weights*: Connect all nodes to each other within clusters and set edge weights to 1 (black solid edges). Connect the inter cluster nodes in a pairwise fashion and set weights to 0.1 (gray solid edges). Finally, connect the outlier node with the clusters and set edge weights very close to $\frac{0.1}{N}$ (gray dashed edges), see Lemma 8.2 in the Appendix for further explanation.

3. *Indexing*: Nodes on the left are indexed by odd numbers and nodes on the right even. The first and the second node correspond to the two outlier nodes respectively.

4. *Graph Laplacian*: The graph Laplacian is taken to be $\frac{1}{\rho_L}L$, where L is the unnormalized graph Laplacian $D - W$.

We choose our initialization by thresholding the second eigenvector, namely, $u^0 = Sign(\phi^2)$. The key property of the graph lies in its second eigenvector, the computation of which could be found in 8.2 in the Appendix :

PROPOSITION 4.5. *Under the setup above, the second eigenvector of the graph*

Laplacian is

$$\phi^2 = \left(\frac{1}{2}, -\frac{1}{2}, \frac{1}{2\sqrt{N}}, -\frac{1}{2\sqrt{N}}, \dots, \frac{1}{2\sqrt{N}}, -\frac{1}{2\sqrt{N}}\right)^t.$$

Moreover, projection of u^0 *onto the first two eigenvectors satisfies* $P_2(u^0) = C\sqrt{N}\phi^2$*, where* C *is approximately* 0.5*.*

Next, we can simply use the formula for the second eigenvector to show that the scheme diverges for $dt = O(N^{-\alpha})$. The idea is that after the first iteration, the values of u^1 on the outlier nodes are high enough such that the scheme diverges. Below is an outline of the proof.

Proof. [Proposition 4.4] We run through the scheme with $u^0 = Sgn(\phi^2)$, $dt = \delta N^{-\alpha}$.

(Step1) We compute u^1. Since u^0 is ±1 valued, $v^0 = u^0$. By boundedness of eigenvalues, $u^1 = C_0\sqrt{N}\phi^2 = O(\sqrt{N})\phi^2$, since C_0 is bounded from below with respect to N.

(Step 2) We compute u^2. Note since $u^1(1) = -u^1(2) = O(\sqrt{N})$, and $dt \leq N^{-\alpha}$, we have $|v^1_1| = |(1-dt)u^1_1 + dt(u^1_1)^3| = O(\sqrt{N}) + O(N^{3/2}/N^{-\alpha}) = O(N^{\theta/2})$, where $\theta > 1$. Similarly, $v^1(j) = O(1)$ for $j \neq 2$. Moreover, by symmetry, $u^1(2k) = -u^1(2k+1)$, and hence we have $v^1 = O(N^{\theta/2})\phi^2 + O(N^{\frac{\theta-1}{2}})$. By calculating projections onto ϕ^2, $u^2 = O(N^{\theta/2})\phi^2 + O(N^{\frac{\theta-1}{2}})$.

(Step 3) Inductively, we can show that $u^k = O(N^{\frac{\theta^k-1}{2}})\phi^2 + O(N^{\frac{\theta^{k-1}-1}{2}})$. And thus letting $k \to \infty$, $u^k \to \infty$. □

Whether the theoretical worst case bound is attained if we project to more than two eigenvectors is not proved here and could be done in future work. However, due to level of freedom in constructing such graphs, the thought is that there are more complicated examples such that the bound is attained for truncation level $m > 2$.

4.3. Heuristic Explanation for Good Typical Behavior. Despite the pathological behavior of the example given above, the timestep for spectral truncation does not depend badly on the size N in practice. In this section, we attempt to give a heuristic explanation of this from two viewpoints.

The first view is to analyze the projection operator P_m in the L^∞ norm. The reason why the maximum principle fails is because P_m is expansive in the L^∞ norm. Namely, for some vector $\|v\|_\infty \leq 1$, we have $\|P_m(v)\|_\infty = O(\sqrt{N})$ in the worst case. However, an easy analysis shows the probability of attaining such an $O(\sqrt{N})$ bound decays exponentially as N grows large, as shown in a simplified analysis in Proposition 8.3 of the Appendix. Thus in practice, it is very rare that adding P_m would violate the maximum principle "too much".

The second view is to restrict our attention to data that come from a random sample. Namely, we assume our data points x^i are sampled i.i.d. from a probability distribution p, and that the graph Laplacian is computed from the Euclidean distance $\|x^i - x^j\|$. In [39], it is proven that under very general assumptions, the discrete eigenfunctions, eigenvalues converges to continuous limits almost surely. Moreover, the projection operators P_k converges compactly almost surely to their continuous limits. Moreover, results for continuous limits of graph-cut problems can be found in [34]. Under this set up, we can define the Allen-Cahn scheme on the continuous domain and discuss its properties on suitable function spaces. The spectral truncated scheme *still* would not satisfy the maximum principle, but at least it evolves in a sample-size independent fashion. Of course a rigorous proof would require heavy functional analysis.

5. Results for Multiclass Classification. The previous analysis can be carried over in a straight forward fashion to the multiclass Ginzburg-Landau.

Multiclass diffuse interface algorithm on graphs can be found in [20, 24, 28]. In most of the algorithms, the labels are vectorized into separate coordinates. To be more precise, given K the number of classes, and N the number of nodes on the graph, we define an $N \times K$ matrix $\boldsymbol{u}$, where each entry $\boldsymbol{u_{ij}}$ stands for the "score" of the *ith* node belonging to the *jth* class. Often one would project $\boldsymbol{u}$ onto the Gibbs simplex $\mathcal{G} = \{\sum x_i = 1 | x_i \geq 0\}$ [21] to make $\boldsymbol{u_{ij}}$ into a probability distribution.

The Ginzburg Landau functional for multiclass is defined as

$$GL(\boldsymbol{u}) = \frac{\epsilon}{2} tr(\boldsymbol{u} L \boldsymbol{u}) + \frac{1}{2\epsilon} \sum_{i=1}^{N} W(\boldsymbol{u_i}). \tag{5.1}$$

Here, $\boldsymbol{u_i}$ stands for the i-th row of $\boldsymbol{u}$, and W should be a "multi-well" function that is analogous to the double-well in the binary case. The function should have local minima near the unit vectors $e_k = (0, 0, \dots, 1, \dots, 0)^t$, and grows fast when u is far from the origin. We use the L^2 double well, namely,

$$W(\boldsymbol{u_i}) = (\Pi_{k=1}^{K} \|\boldsymbol{u_i} - e_k\|_2^2). \tag{5.2}$$

In [21], a different double well is used where L^1 norms are taken instead of L^2. The paper claimed that L^2 double well suffers from the problem that the function value of W in the center of the Gibbs simplex is small. This problem could be alleviated if we rescale distance by a suitable function $\rho(x)$. Namely, replacing $\|\boldsymbol{u_i} - e_k\|_2^2$ by $\rho(\|\boldsymbol{u_i} - e_k\|_2^2)$. Moreover, when k is reasonably small, even such adjustments are unnecessary. However, choosing the L^2 well comes with the bonus advantage that the problem is smooth. This gives better convergence guarantees as well as makes the problem easier to compute numerically. For example, the L^2 double well does not require projection onto the Gibbs simplex $\mathcal{G}$ in every iteration as in [21].

We minimize GL using the forward-backward method as in (2.4).

$$\begin{cases} \boldsymbol{v^k} = \boldsymbol{u^k} - dt * \dfrac{1}{2\epsilon} \displaystyle\sum_i \nabla W(\boldsymbol{u_i^k}), \\ \boldsymbol{u^{k+1}} = -dt * (\epsilon L \boldsymbol{u^{k+1}}) + \boldsymbol{v^k}. \end{cases} \tag{5.3}$$

Since the diffusion step is done columns-wise, the maximum principle carries over naturally. Namely, we have

PROPOSITION 5.1 (Maximum Principle Multiclass). *Let $\boldsymbol{u}, \boldsymbol{v}$ be $K \times N$ matrices. Define*

$$\boldsymbol{u} = -dt * (\epsilon L \boldsymbol{u}) + \boldsymbol{v}, \tag{5.4}$$

then $\|u_j\|_\infty \leq \|v_j\|_\infty$, where u_j, v_j are the jth column of u, v respectively.

With the exact same reasoning as in the binary case, we need a range of stepsize dt for which the forward gradient step $\mathcal{F}_t$ of the "multi-well" maps $[-R, R]^{N \times K}$ onto itself, as the next lemma shows.

LEMMA 5.2. *Define $\mathcal{F}_t : \boldsymbol{u_i} \mapsto \boldsymbol{u_i} - dt * \frac{1}{2\epsilon} \nabla W(\boldsymbol{u_i})$. Then $\exists R(K)$ and $\exists c(R, K)$ independent of N such that for $dt < c(R, K)$, $\mathcal{F}_t([-R, R]^K) \subset [-R, R]^K$.*

Proof. Since the double well W does not depend on N, the constants R and dt are naturally independent of N if we prove its existence.

We define a new map ϕ_{dt} to be $\phi_{dt}(R) = \sup\{\|F_{dt}(u)\|_\infty, u \in \partial[-R,R]^K\}$. It can be shown that ϕ_{dt} is continuous. Since the double well W is nearly quadratic when R is large, we have that $\exists R_1$ such that $\nabla W|_{\partial[-R_2,R_2]^K}$ points inward of the box $[-R_2, R_2]^K$, for all $R_2 \geq R_1$. Thus we can find c dependent on R_2 such that $\phi_{dt}(R_2) \leq R_2, \forall R_2 \geq R_1, dt < c$. Take $R = \max \phi_c([0, R_1])$, by shrinking c if necessary, we have $\max \phi_c([0,R]) \leq R$, and thus $\mathcal{F}_t([-R,R]^K) \subset [-R,R]^K, \forall dt < c$. □

The proof works for any function that acts independently on each component $\boldsymbol{u_i}$ and has fast growth towards infinity. The estimates here are not as precise as the 0.5ϵ bound in the binary case, since an explicit calculation will be a rather complicated formulae that involves K. However, in practice, the stepsize restriction is also comparable to 0.5ϵ, at least when the number of clusters K is moderate.

Additional fidelity terms and alternative graph Laplacian could be handled the same way as in the binary case. Hence we have,

THEOREM 5.3 (Convergence). *The multiclass graph Allen-Cahn scheme, with or without fidelity, is convergent for stepsize $dt < c\epsilon$.*

6. Numerical Results.

In this section, we construct various numerical experiments of increasingly larger scales. This helps demonstrate our theory, and also have some implication on the real world performance of the schemes.

6.1. Two Moons.

The two moons data was used by Buhler et al [10] in exploring spectral clustering with p-Laplacians. It is constructed from sampling from two half circles of radius one on $\mathbb{R}^2$, centered at (0,0) and (1,0.5). Gaussian noise of standard deviation 0.02 in $\mathbb{R}^{100}$ is then added to each of the points. The weight matrix is constructed using Zelnik-Manor and Perona's procedure [41]. Namely, set $w_{ij} = e^{-d(i,j)/\sqrt{\tau_i \tau_j}}$, where τ_i is the Mth closest distance to i. W is further symmetrized by taking the max between two symmetric entries.

Fig.6.1 is an illustration of the data set of three different sizes being segmented perfectly under a uniform stepsize. A zero mass constraint is used instead of fidelity points, and random initialization is chosen. The parameters for the experiment is $dt = 0.5, \epsilon = 1$, which is exactly the tight bound.

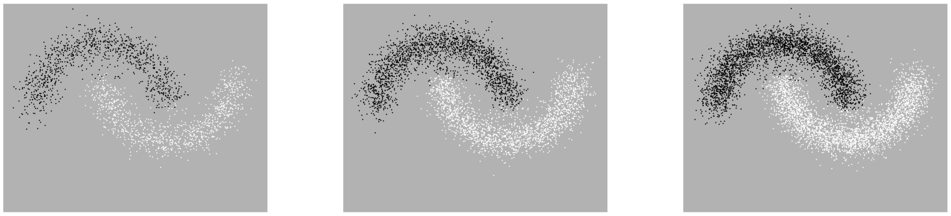


Fig. 6.1: Segmentation results under the same stepsize for Two Moons with sample sizes 1000, 2000, 3000 respectively.

To test the theory in a more rigours manner, we compute several "maximum stepsizes" that ensures some criterion (e.g. bounded after 500 iterations, etc.), and compare this with the stepsize predicted by the theory. Bisection with 1e-5 accuracy

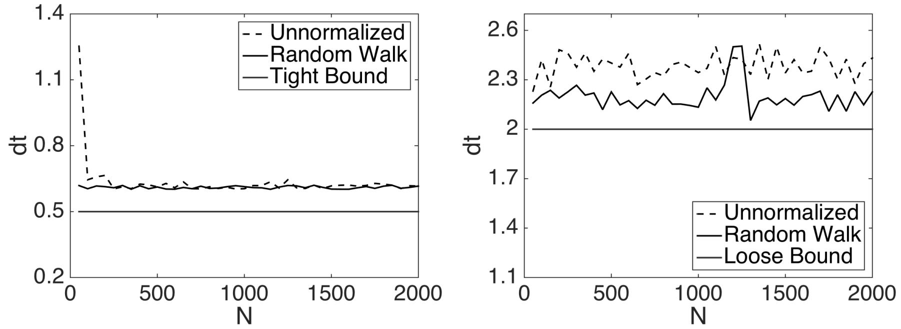


Fig. 6.2: Two Moons Segmentation Problem. Left: Maximum stepsize satisfying $\|u^k\|_\infty \leq 1$. Right: Left: Maximum stepsize satisfying $\|u^k\|_\infty \leq 10$. N is the number of nodes.

is used to determine the maximum stepsize that satisfies the criterion given.

Fig 6.2 plots the maximum stepsize for the scheme (1.9) to be bounded by 1.0005, 10 respectively. Random $-1, 1$ initial conditions are chosen. No fidelity terms are added and the diffuse parameter $\epsilon = 1$. We also compute results for the random walk Laplacian and the unnormalized Laplacian as comparison. The actual results are independent of graph size, and also match the tight and loose bound nicely.

Since we are interested in convergence stepsize, we switch our criterion from boundedness to convergence, namely, we compute the stepsizes for which the scheme has iterative difference less than 1e-4 in 1000 iterations. ϵ is still chosen to be 1.

Fig.6.3 (left) plots the limit convergence stepsize for the scheme with the three different Laplacians. As we can see, the typical limit stepsize is between the tight and loose bound. Fig.6.3 (right) fixes $N = 2000$ and varies ϵ to plot the relation between dt and ϵ. They are almost linear as predicted by the 0.5ϵ bound.

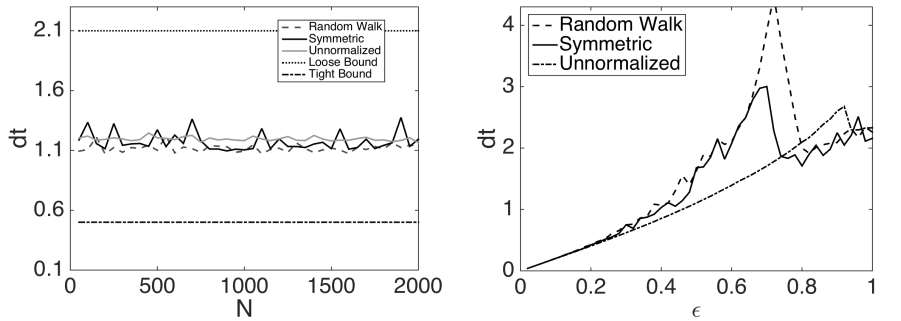


Fig. 6.3: Two Moons Segmentation Problem. Left: Maximum stepsize for convergence, fixing $\epsilon = 1$ varying N. Right: Maximum stepsize for convergence, fixing $N = 2000$ varying ϵ, ϵ is the interface scale parameter. N is the number of nodes on the graph.

Fig.6.4 (left) plots the scheme with spectral truncation. The results are compared with the full scheme, and are roughly in the same range. Fig.6.4 (right) plots the

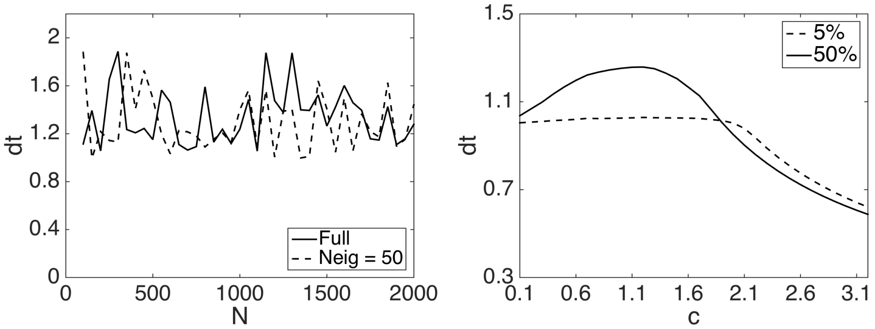


Fig. 6.4: Two Moons Segmentation Problem. Left: Maximum convergence stepsize comparing spectral truncation vs full scheme. Right: Maximum convergence stepsize for scheme with fidelity.N the number of nodes. c is the fidelity strength.

effects of adding a quadratic fidelity term with power c while keeping $\epsilon = 1$ fixed. As we can see from the result, the fidelity term does constitute an additional restriction when c is large. However, stepsizes remain roughly the same for small c. It is hard to analyze the exact effect when c and ϵ are comparable.

6.2. Worst Case Graph. Despite the good practical behavior of spectral truncation, this experiment shown in Fig.6.5 is a realization of the worst case stepsize restriction for spectral truncation. The plot in log-log axis shows the convergent dt vs the size of the problem N. The scheme is initialized by thresholding the 2nd eigenvector. The slope of the descent matches the theoretical $k = -1$ line almost exactly, proving the optimality of the theoretical result.

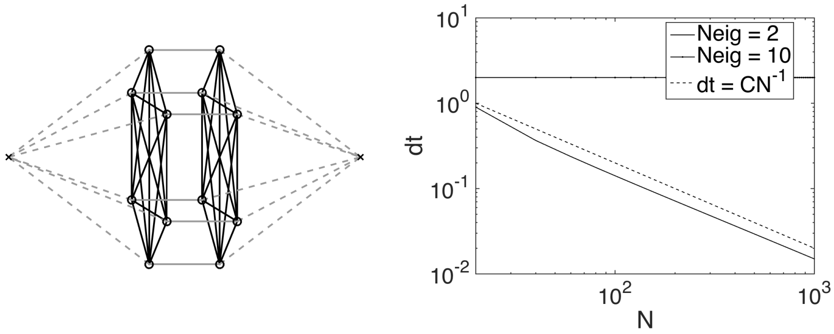


Fig. 6.5: Left: Cartoon figure of the worst case graph. Right: Log plot of maximum stepsize for $\{u^k\}$ to be bounded. N the number of nodes on the graph.

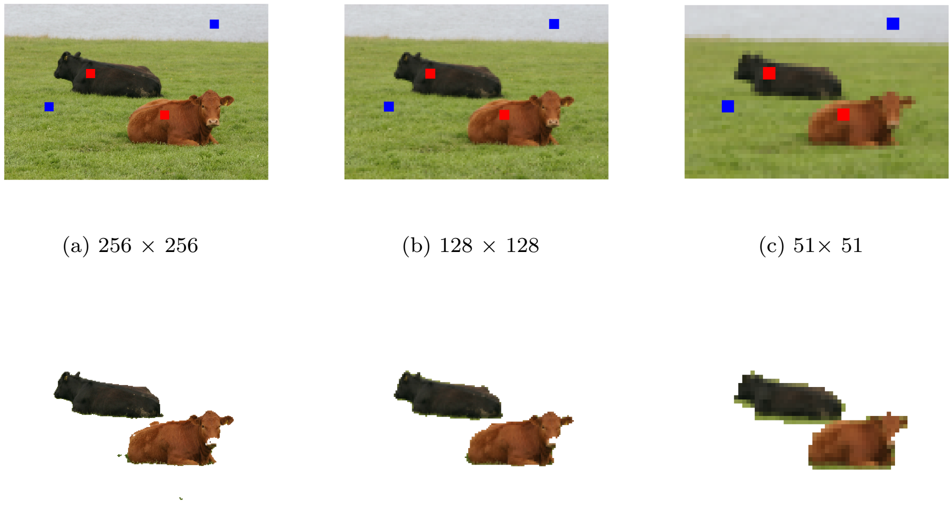
6.3. Two Cows. The point of this experiment is to test the effects of Nystrom sampling on the stepsize and overall performance of the algorithm. The images of the two cows are from the Microsoft Database. For large dense graphs such as nonlocal graphs from images, it is often impractical to compute the entire graph. Nystrom sampling is a technique used to approximate eigenvectors without explicitly computing the graph Laplacian [3, 18, 19].

From the original 312×280 image, we generate 10 images with successively lower resolution of $(312/k) \times (280/k)$, $k = 1, \ldots 10$. A non-local graph constructed from feature windows of size 7×7 is used, and weights are constructed by the standard Gaussian Kernel $w_{ij} = e^{-d_{ij}/\sigma^2}$. The eigenvectors are constructed by using Nystrom extension, the details of which could be found in [5].

Nystrom extension produces an orthogonal set of vectors that approximates the true eigenvectors by subsampling from the original graph. The following examples show that this imprecision does not cause numerical instability.

Fig.6.6 illustrates three images with 1,1/2,1/5 times original resolution being segmented under the uniform condition $dt = 2$, $\epsilon = 4$. The blue and red box corresponds to fidelity points of the two classes, the constant in front of the fidelity are $c_1 = 1$ and $c_2 = 0.4$ for the cows and the background respectively.

Fig.6.7 is a profile of N vs dt. To ensure segmentation quality, smaller epsilon had to be chosen for images of lower resolution, and the final result is displayed in terms of the dt/ϵ ratio.



(a) 256 × 256

(b) 128 × 128

(c) 51× 51

(d) Segmentation Result 1

(e) Segmentation Result 2

(f) Segmentation Result 3

Fig. 6.6: Images of different resolution segmented under the same stepsize

6.4. MNIST. This experiment is used to demonstrate the case of multiclass clustering by the L^2 multiclass Ginzburg-Landau functional.

The MNIST database [26] is a data set of 70000 28×28 images of handwritten digits from 0-9. The graph is constructed by first doing a PCA dimension reduction and again using the same Zelnik-Manor and Perona's procedure with 50 nearest neighbors.

For our purpose, we focus on clusters of size three. Table 6.1 shows the limit stepsizes of various tuples, and the error rate when segmented under a uniform stepsize.

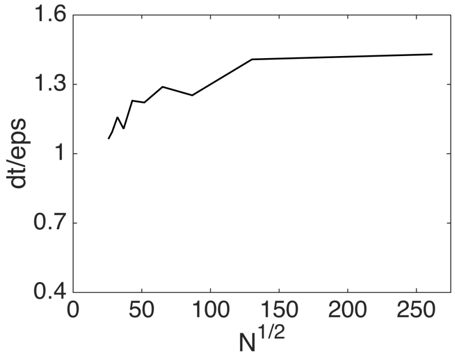


Fig. 6.7: Maximum Convergence Stepsize for Two Cows a Series of Different Resolution. N is the number of nodes in the graph, which equals $A \times B$ with A, B the height and width of an image.

5% fidelity points are used, and $\epsilon = 1$. The scheme is projected onto the first 100 eigenvectors. It is shown here that they are still segmented around the same stepsize.

Tuples	{4,6,7}	{3,5,8}	{1,0,9}	{0,6,1}	{2,7,1}
Max dt	0.5823	0.5914	0.5716	0.5701	0.5755
Correct (dt=0.5)	97.98%	97.58%	96.00%	96.36%	98.22%

Table 6.1: Clustering results of MNIST. For each digit, $N \approx 6000$. First Row: triplets of digits to be classified. Second Row: Maximum stepsize for convergence. Third Row: Error rate with a fixed dt that is close to the maximum stepsize.

7. Discussion. In summary, we show that the semi-implicit scheme for solving the graph Allen-Cahn equation converges to a stationary point under a graph-independent stepsize $dt = 0.5\epsilon$. The proof combines ideas form classical numerical analysis and also convex analysis. We then analyze the same convergence stepsize problem for the scheme under spectral truncation. We show that unlike the previous case, a graph-independent stepsize bound that works on all graphs is no longer possible. This is because maximum principle no longer holds under spectral truncation. A new bound $dt = O(N^{-1})$ is obtained and is shown to be sharp in the worst case. Some heuristics were provided to explain the discrepancy between the worst case performance and the good average case behavior when applying spectral truncation. We then present a natural extension of the analysis to multi-class classification. We finally conduct a variety of numerical experiments on various datasets to demonstrate how the theory matches practical performance.

There are still some very interesting problems left to be explored. One important problem is why so few eigenvectors are needed during spectral truncation? Is this unique for classification tasks, and how can this be quantified in a more theoretical framework? Another problem is the relationship between the stepsize and final error rate. As the problem is non-convex, converging to a sub-optimal stationary point is a possibility. So far this analysis does not attempt to characterize the quality of the final converged solution, but experiments have shown that the error rates do differ under different stepsize.

8. Appendix. LEMMA 8.1 (Norm Conversions). *Let $1 \leq p < q \leq +\infty$. Then the formula below explicitly states the equivalence between norms:*

$$\|u\|_q \leq \|u\|_p \leq \|u\|_q N^{1/p-1/q}.$$

Proof. The right hand side is by a generalization of Holder's inequality. The left hand side is by simple polynomial expansion. □

LEMMA 8.2 (Computation of Second Eigenvector of Graph 4.1). *The second eigenvector of the graph in Fig. 4.1 is*

$$\phi^2 = \left(\frac{1}{2}, -\frac{1}{2}, \frac{1}{2\sqrt{N}}, -\frac{1}{2\sqrt{N}}, \dots, \frac{1}{2\sqrt{N}}, -\frac{1}{2\sqrt{N}}\right)^t.$$

Proof. We set the gray solid edges have weights α, and the gray dashed edges β/n. Recall the variational formulation of the second eigenvector

$$\operatorname*{argmin}_u Dir(u) \quad s.t. \quad \langle u, e^1 \rangle = 0, \|u\|_2 = 1.$$

Note that by symmetry, we can assume $\phi^2 = (a, -a, b, -b, \dots, b, -b)^t$. Under this parameterization, we have that the Dirichlet energy is

$$Dir(\alpha, \beta) = \frac{\alpha}{n}(b-a)^2 \times n + \beta(2b)^2 n. \tag{8.1}$$

Hence by computing the Lagrange multipliers, we have

$$\begin{cases} nkb = & 2\gamma nb - (b-a), \\ ka = & a - b, \end{cases} \tag{8.2}$$

where $\gamma = \frac{\beta}{\alpha}$, and k is the lagrange multiplier. The equation for k is

$$k^2 - (\frac{1}{n} + 2\gamma + 1)k + 2\gamma = 0. \tag{8.3}$$

Setting $2\gamma = 1+\theta$, and computing the roots k, we have $k = 1-(\sqrt{\frac{1}{n} - \theta + (\frac{1}{n} + \theta)^2/4 - \theta^2} - \frac{\theta}{2} - \frac{1}{2n})$. Setting $\theta = 0$ gives $k = 1 - \frac{1}{\sqrt{n}} - \frac{1}{2n}$. However, we need $k = 1 - \frac{1}{\sqrt{n}}$ to yield our desired eigen-vector. This can be done by setting the correction term $\theta = o(\frac{1}{n})$. □

PROPOSITION 8.3. *Define the set*

$$M = \{\, u \in \mathbb{R}^N \mid \|u\|_\infty \leq 1, \max_{P_m} \|P_m u\|_\infty \geq C\sqrt{N} \,\},$$

where P_m is any projection operator onto a subspace, and $0 < C < 1$. Then the volume(with respect to the standard L^2 metric in R^N) of the set M decreases exponentially with respect to the number of dimensions N.

The proposition shows that if u were sampled uniformly from a unit cube, then the probability of some projection P_m expanding the max norm by a factor of $O(\sqrt{N})$ is exponentially decreasing in N.

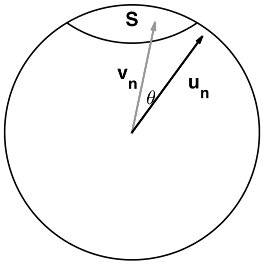


Fig. 8.1: Illustration of Proposition 8.3. S is one of the "caps" that v_n resides in. u_n and v_n have angle less than θ.

Proof. Let $u \in M$. Then by definition of the set M, $\exists$ some projection P_m such that $\|P_m u\|_\infty \geq C\sqrt{N}$. Define $v := P_m u$ and $v_n := \frac{v}{\|v\|_2}$. Define $u_n := \frac{u}{\|u\|_2}$. Since v_n is the projected direction of u, $P_m u = \langle u, v_n \rangle v_n$. Then we have

$$C\sqrt{N} \leq \|P_m u\|_\infty = \langle u, v_n \rangle \|v_n\|_\infty = \|u\|_2 \|v_n\|_\infty \langle u_n, v_n \rangle.$$

Since $\|u\|_2 \leq \sqrt{N}$, we have

$$\|v_n\|_\infty \langle u_n, v_n \rangle \geq C. \tag{8.4}$$

Since $\langle u_n, v_n \rangle \leq 1$, the projected direction v_n must be in the set $S = \{v \mid \|v\|_2 = 1, \|v\|_\infty \geq C\}$. However, the set S contains the N "caps" of a unit sphere (see Fig.8.1), and hence is exponentially decreasing in volume with respect to the sphere. On the other hand, since $\|v_n\|_\infty \leq 1$, by (8.4) we have $\langle u_n, v_n \rangle \geq C$, and thus u lies in a cone $K(v_n)$ with angle $cos(\theta) \geq C$. Hence $u \in K_v + N$, and since cones K_v have volume exponentially decreasing with respect to N as well, we have $Vol(M)$ is exponentially decreasing with respect to N.
□

REFERENCES

[1] Christopher R Anderson. A Rayleigh–Chebyshev procedure for finding the smallest eigenvalues and associated eigenvectors of large sparse Hermitian matrices. *Journal of Computational Physics*, 229(19):7477–7487, 2010.

[2] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

[3] Serge Belongie, Charless Fowlkes, Fan Chung, and Jitendra Malik. Spectral partitioning with indefinite kernels using the Nystrom extension. In *Computer Vision ECCV 2002*, pages 531–542. Springer, 2002.

[4] Andrea L Bertozzi, Selim Esedoglu, and Alan Gillette. Inpainting of binary images using the Cahn-Hilliard equation. *IEEE Transactions on image processing*, 16(1):285–291, 2007.

[5] Andrea L Bertozzi and Arjuna Flenner. Diffuse interface models on graphs for classification of high dimensional data. *Multiscale Modeling & Simulation*, 10(3):1090–1118, 2012.

[6] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2009.

[7] Xavier Bresson, Thomas Laurent, David Uminsky, and James V Brecht. Convergence and energy landscape for Cheeger cut clustering. In *Advances in Neural Information Processing Systems*, pages 1385–1393, 2012.

[8] Xavier Bresson and Arthur D Szlam. Total variation, Cheeger cuts. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 1039–1046, 2010.

[9] Xavier Bresson, Xue-Cheng Tai, Tony F Chan, and Arthur Szlam. Multi-class transductive learning based on $\ell 1$ relaxations of Cheeger cut and Mumford-Shah-Potts model. *Journal of mathematical imaging and vision*, 49(1):191–201, 2014.

[10] Thomas Bühler and Matthias Hein. Spectral clustering based on the graph p-Laplacian. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 81–88. ACM, 2009.

[11] Antonin Chambolle and Jérôme Darbon. On total variation minimization and surface evolution using parametric maximum flows. *International journal of computer vision*, 84(3):288–307, 2009.

[12] Fan Chung and Wenbo Zhao. PageRank and random walks on graphs. In *Fete of combinatorics and computer science*, pages 43–62. Springer, 2010.

[13] Philippe G Ciarlet. Discrete maximum principle for finite-difference operators. *Aequationes mathematicae*, 4(3):338–352, 1970.

[14] Joseph B Collins and Herbert Levine. Diffuse interface model of diffusion-limited crystal growth. *Physical Review B*, 31(9):6119, 1985.

[15] Chris HQ Ding, Xiaofeng He, Hongyuan Zha, Ming Gu, and Horst D Simon. A min-max cut algorithm for graph partitioning and data clustering. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 107–114. IEEE, 2001.

[16] Heike Emmerich. *The diffuse interface approach in materials science: thermodynamic concepts and applications of phase-field models*, volume 73. Springer Science & Business Media, 2003.

[17] David J Eyre. An unconditionally stable one-step scheme for gradient systems. *Unpublished article*, 1998.

[18] Charless Fowlkes, Serge Belongie, Fan Chung, and Jitendra Malik. Spectral grouping using the Nystrom method. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(2):214–225, 2004.

[19] Charless Fowlkes, Serge Belongie, and Jitendra Malik. Efficient spatiotemporal grouping using the Nystrom method. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–231. IEEE, 2001.

[20] Cristina Garcia-Cardona, Arjuna Flenner, and Allon G Percus. Multiclass semi-supervised learning on graphs using Ginzburg-Landau functional minimization. In *Pattern Recognition Applications and Methods*, pages 119–135. Springer, 2015.

[21] Cristina Garcia-Cardona, Ekaterina Merkurjev, Andrea L Bertozzi, Arjuna Flenner, and Allon G Percus. Multiclass data segmentation using diffuse interface methods on graphs. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(8):1600–1613, 2014.

[22] Stephen Guattery and Gary L Miller. On the quality of spectral separators. *SIAM Journal on Matrix Analysis and Applications*, 19(3):701–719, 1998.

[23] Huiyi Hu, Thomas Laurent, Mason A Porter, and Andrea L Bertozzi. A method based on total variation for network modularity optimization using the MBO scheme. *SIAM Journal on Applied Mathematics*, 73(6):2224–2246, 2013.

[24] Huiyi Hu, Justin Sunu, and Andrea L Bertozzi. Multi-class graph Mumford-Shah model for plume detection using the MBO scheme. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 209–222. Springer, 2015.

[25] Tom Ilmanen. Convergence of the Allen-Cahn equation to Brakke's motion by mean curvature. *J. Differential Geom.*, 38(2):417–461, 1993.

[26] Yann LeCun and Corinna Cortes. The MNIST database of handwritten digits http://yann.lecun.com/exdb/mnist/, 1998.

[27] Ekaterina Merkurjev, Egil Bae, Andrea L Bertozzi, and Xue-Cheng Tai. Global binary optimization on graphs for classification of high-dimensional data. *Journal of Mathematical Imaging and Vision*, 52(3):414–435, 2015.

[28] Ekaterina Merkurjev, Tijana Kostic, and Andrea L Bertozzi. An MBO scheme on graphs for classification and image processing. *SIAM Journal on Imaging Sciences*, 6(4):1903–1930, 2013.

[29] Barry Merriman, James K Bence, and Stanley J Osher. Motion of multiple junctions: A level set approach. *Journal of Computational Physics*, 112(2):334–363, 1994.

[30] Luciano Modica and Stefano Mortola. Un esempio di γ-convergenza. *Boll. Un. Mat. Ital. B (5)*, 14(1):285–299, 1977.

[31] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.

[32] Suvrit Sra. Scalable nonconvex inexact proximal splitting. In *Advances in Neural Information Processing Systems*, pages 530–538, 2012.

[33] John C Strikwerda. *Finite difference schemes and partial differential equations*. Siam, 2004.

[34] Nicolas Garcia Trillos, Dejan Slepcev, James von Brecht, Thomas Laurent, and Xavier Bresson. Consistency of Cheeger and Ratio graph cuts. *arXiv preprint arXiv:1411.6590*, 2014.

[35] Yves Van Gennip, Andrea L Bertozzi, et al. Gamma -convergence of graph Ginzburg-Landau functionals. *Advances in Differential Equations*, 17(11/12):1115–1180, 2012.

[36] Yves van Gennip, Nestor Guillen, Braxton Osting, and Andrea L Bertozzi. Mean curvature, threshold dynamics, and phase field theory on finite graphs. *Milan Journal of Mathematics*, 82(1):3–65, 2014.

[37] Benjamin P Vollmayr-Lee and Andrew D Rutenberg. Fast and accurate coarsening simulation with an unconditionally stable time step. *Physical Review E*, 68(6):066703, 2003.

[38] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[39] Ulrike Von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, pages 555–586, 2008.

[40] Alan L Yuille, Anand Rangarajan, and AL Yuille. The concave-convex procedure (CCCP). *Advances in neural information processing systems*, 2:1033–1040, 2002.

[41] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *Advances in neural information processing systems*, pages 1601–1608, 2004.